

VARIABLE SELECTION IN ROBUST JOINT MEAN AND COVARIANCE MODEL FOR LONGITUDINAL DATA ANALYSIS

Xueying Zheng¹, Wing Kam Fung² and Zhongyi Zhu¹

¹*Fudan University* and ²*The University of Hong Kong*

Abstract: In longitudinal data analysis, a correct specification of the within-subject covariance matrix cultivates an efficient estimation for mean regression coefficients. In this article, we consider robust variable selection method in a joint mean and covariance model. We propose a set of penalized robust generalized estimating equations to simultaneously estimate the mean regression coefficients, the generalized autoregressive coefficients, and innovation variances introduced by the modified Cholesky decomposition. The set of estimating equations select important covariate variables in both mean and covariance models together with the estimating procedure. Under some regularity conditions, we develop the oracle property of the proposed robust variable selection method. Finally, a simulation study and a detailed data analysis are carried out to assess and illustrate the small sample performance; they show that the proposed method performs favorably by combining the robustifying and penalized estimating techniques together in the joint mean and covariance model.

Key words and phrases: Covariance matrix, penalized generalized estimating equation, longitudinal data, modified cholesky decomposition, robustness, variable selection.

1. Introduction

Longitudinal data arise more and more frequently in a variety of scientific domains that seek insightful and comprehensive research in a branch of statistical modeling. Different from other types of data, we often assume independence among distinct subjects but dependence within each subject; within-subject correlation raises a fundamental challenge for the analysis of longitudinal data. Liang and Zeger (1986), a milestone in the development of methodology for longitudinal data analysis, proposed generalized estimating equations (GEE) for estimation of generalized linear regression coefficients. The main advantage of their method is that even when the within-subject correlation is treated as a nuisance parameter with an assumed parsimonious structure, GEE still brings about a consistent estimator for the mean regression model. Subsequential, Qu, Lindsay, and Li (2000) used the quadratic inference function (QIF) to enhance

the efficiency by considering the structure of the covariance matrix. Taking robustification into account, He, Fung, and Zhu (2005) proposed the robust GEE method to prevent the unexpected influence from outliers in a longitudinal data set.

Ignoring the within-subject correlation can result in an inefficient estimator of a regression model. In practice, the within-subject covariance structure itself may be of scientific interest. Relevant topics here include component analysis and factor analysis in multivariate statistical problems. Recent research on the estimation of the covariance matrix includes, but is not limited to, Rothman, Levina, and Zhu (2009), El Karoui (2008) and Bickel and Levina (2008).

Similar to the mean regression, covariances may be dependent on various explanatory variables. Pourahmadi (1999, 2000) proposed a joint mean and covariance regression model by decomposing the covariance matrix employing generalized autoregressive coefficients and innovation variances. Ye and Pan (2006) extended the joint model under the framework of generalized estimating equations which required no assumptions on the distribution of the data. By introducing generalized autoregressive coefficients and innovation variances, their joint model eliminated the positive definiteness constraint in estimation of the covariance matrix. Instead, three generalized estimating equations were proposed to estimate the covariance matrix and the mean simultaneously.

A number of developments have been followed, see Fan, Huang, and Li (2007), Fan and Wu (2008), and Xu and Mackenzie (2012). Leng, Zhang, and Pan (2010) generalized Ye and Pan's model to a semiparametric joint mean and covariance model. Mao, Fung, and Zhu (2011) extended Leng, Zhang, and Pan (2010)'s work further to a generalized partially linear varying coefficient model. Zheng, Fung, and Zhu (2013) extended the robust estimating equation in He, Fung, and Zhu (2005) to the joint mean and covariance model by creating three robust estimating equations to mitigate the effect of outliers in both mean and covariance estimation.

Variable selection is a technique for selecting a subset of relevant covariates in constructing reliable statistical models. Many variable selection methods are based on the penalized likelihood or penalized estimating equations. Commonly used penalties include LASSO, ALASSO (adaptive lasso in Zou (2006)) SCAD, and Hard penalties. In longitudinal data analysis, Fu (2003) proposed the penalized generalized estimating equation with LASSO penalty. Other variable selection criteria include AIC and C_p , extended by Pan (2001) and Cantoni, Flemming, and Ronchetti (2005), respectively, to the case of longitudinal data under the framework of GEE.

In contrast, little work has been done on covariance variable selection or identification. Jeng and Daye (2011) noticed that to curve the sparsity in the

covariance matrix can improve the efficiency on mean estimation, especially in high dimensional problems, though their main objective was still the mean estimation. Within the framework of the joint mean and covariance model, Huang et al. (2006) proposed covariance selection and estimation via penalized normal likelihood. Kou and Pan (2011) proposed a penalized maximum likelihood method for the joint model that employed variable selection in both models simultaneously, the covariance matrix being treated as of equal importance as the mean.

In this paper, we develop a penalized robust estimating equations based method to select important explanatory variables. Simulation studies have shown that both the classical GEE and the joint mean and covariance model are sensitive to outliers, see details in He, Fung, and Zhu (2005) and Zheng, Fung, and Zhu (2013). Nevertheless, the discussion on robust variable selection methods has been limited. We consider the SCAD and ALASSO penalties, and show the oracle property of the proposed robust variable selection method. In data analysis, the robust variable selection procedure can provide standard errors for mean coefficients estimation, due to sensible covariance matrix modeling and the accommodation of outliers in both subject and observation levels.

The remainder of this article is organized as follows. Section 2 describes the main model and its asymptotic properties. Simulation results are given in Section 3. In Section 4, we apply the proposed method to a hormone data set.

2. Robust Variable Selection in Joint Mean and Covariance Model

2.1. Joint mean and covariance model

Suppose that we have a sample of m subjects. Let $y_i = (y_{i1}, \dots, y_{in_i})^T$ be the n_i repeated observations at time point $t_i = (t_{i1}, \dots, t_{in_i})^T$ of the i th subject. Let $E(y_i) = \mu_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ and $Cov(y_i) = \Sigma_i$ be the $n_i \times 1$ mean vector and $n_i \times n_i$ covariance matrix of y_i , respectively.

To eliminate the constraint of positive definiteness in estimation of the covariance matrix Σ_i , we implement a modified Cholesky decomposition by introducing a unique lower triangular matrix Φ_i with 1's as diagonal entries and a unique diagonal matrix D_i with positive diagonals such that

$$\Phi_i \Sigma_i \Phi_i^T = D_i.$$

Here the lower-diagonal entries of Φ_i are the negatives of the autoregressive coefficients ϕ_{ijk} defined in

$$\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}),$$

the linear least squares predictor of y_{ij} based on its predecessors $y_{i(j-1)}, \dots, y_{i1}$. The diagonal entries σ_{ij}^2 of D_i are the innovation variances $\sigma_{ij}^2 = \text{var}(\varepsilon_{ij})$, where $\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$.

We adopt linear models for the mean, generalized autoregressive parameters, and innovation variances as in Ye and Pan (2006):

$$\mu_{ij} = x_{ij}^T \beta, \phi_{ijk} = g_{ijk}^T \gamma, \log \sigma_{ij}^2 = z_{ij}^T \lambda, \quad (2.1)$$

where x_{ij} , g_{ijk} , and z_{ij} are $p \times 1$, $q \times 1$ and $d \times 1$ vectors of covariates, and β , γ and λ are associated parameters. The covariates g_{ijk} and z_{ij} may contain baseline covariates, time and associated interactions. The log-linear model of the innovation variance follows Cook and Weisberg's (1983) model for the variance. Cook and Weisberg developed a log-linear model that allows for dependence of the variance on an arbitrary set of variables. In practice, the whole set of the covariates is difficult to define. An orthogonal form for a polynomial of time was recommended as the covariate for the autoregressive component by Ye and Pan (2006):

$$g_{ijk} = (1, (t_{ij} - t_{ik}), (t_{ij} - t_{ik})^2, \dots, (t_{ij} - t_{ik})^{q-1})^T.$$

The linear assumption in (2.1) can be replaced by quadratic assumptions, or the use of nonparametric or semiparametric models after the decomposition, see Mao, Fung, and Zhu (2011) and Leng, Zhang, and Pan (2010). To simplify matters we start from the linear assumption.

In the models at (2.1), estimation for generalized autoregressive coefficients and innovation variances are treated as important as the estimation for the mean. Let $\theta = (\theta_1, \dots, \theta_s)^T = (\beta_1, \dots, \beta_p; \gamma_1, \dots, \gamma_q; \lambda_1, \dots, \lambda_d)^T$, where $s = p + q + d$. To select important subsets of the covariates, we assume that all interesting explanatory variables, together with their interactions, are involved. By using the same λ , the proposed method is applicable for correlated data as long as the correlated (or longitudinal) measurements have similar correlation structure between clusters.

2.2. Penalized robust generalized estimating equations

We propose a penalized robustified generalized estimating equations

$$U(\theta) = ([U_1(\beta)]^T, [U_2(\gamma)]^T, [U_3(\lambda)]^T)^T,$$

for the mean, generalized autoregressive parameters, and innovation variances, respectively:

$$U_1(\beta) = \sum_{i=1}^m X_i^T (V_i^\beta)^{-1} h_i^\beta(\mu_i(\beta)) - m q_{\tau(1)}(|\beta|) \text{sgn}(\beta) = 0, \quad (2.2)$$

$$U_2(\gamma) = \sum_{i=1}^m T_i^T (V_i^\gamma)^{-1} h_i^\gamma(\hat{r}_i(\gamma)) - m q_{\tau(2)}(|\gamma|) \text{sgn}(\gamma) = 0, \quad (2.3)$$

$$U_3(\lambda) = \sum_{i=1}^m Z_i^T D_i (V_i^\lambda)^{-1} h_i^\lambda(\sigma_i^2(\lambda)) - m q_{\tau(3)}(|\lambda|) \text{sgn}(\lambda) = 0, \quad (2.4)$$

where $h_i^\beta(\mu_i(\beta)) = W_i^\beta[\psi^\beta(\mu_i(\beta)) - C_i^\beta(\mu_i(\beta))]$, $h_i^\gamma(\hat{r}_i(\gamma)) = W_i^\gamma[\psi^\gamma(\hat{r}_i(\gamma)) - C_i^\gamma(\hat{r}_i(\gamma))]$ and $h_i^\lambda(\sigma_i^2(\lambda)) = W_i^\lambda[\psi^\lambda(\sigma_i^2(\lambda)) - C_i^\lambda(\sigma_i^2(\lambda))]$ act as the core of the estimating equations with $X_i = (x_{i1}^T, \dots, x_{in_i}^T)^T$, $Z_i = (z_{i1}^T, \dots, z_{in_i}^T)^T$, $g_{ij} = (g_{ij1}^T, \dots, g_{ij(j-1)}^T)^T$ and $T_i = (g_{i1}^T, \dots, g_{in_i}^T)^T$.

Equations (2.3) and (2.4) are in agreement with (2.2), in which r_i in $U_2(\gamma)$ and ε_i^2 in $U_3(\lambda)$ play roles similar to that of y_i in $U_1(\beta)$, and can be viewed as working responses. Here covariance estimation is treated the same mean estimation in the proposed model, in the spirit of the joint model in Ye and Pan (2006).

We specify items in (2.2)–(2.4). In $U_2(\gamma)$, r_i and \hat{r}_i are $n_i \times 1$ vectors with j th components $r_{ij} = y_{ij} - \mu_{ij}$ and $\hat{r}_{ij} = E(r_{ij}|r_{i1}, \dots, r_{i(j-1)}) = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$, where $\sum_{k=1}^0$ is zero when $j = 1$. In $U_3(\lambda)$, $\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$ and ε_i^2 and σ_i^2 are $n_i \times 1$ vectors with j th components ε_{ij}^2 and σ_{ij}^2 , respectively, where in fact $E(\varepsilon_i^2) = \sigma_i^2$. Moreover, $T_i^T = \partial \hat{r}_i^T / \partial \gamma$ is the $q \times n_i$ matrix with the j th column $\partial \hat{r}_{ij} / \partial \gamma = \sum_{k=1}^{j-1} r_{ik} g_{ijk}$ and $D_i = \text{diag}\{\sigma_{i1}^2, \dots, \sigma_{in_i}^2\}$.

Again at (2.2)–(2.4), $V_i^\beta = A_i^{-1/2} \Sigma_i$, A_i is the diagonal elements of Σ_i , $V_i^\gamma = D_i^{1/2}$, $V_i^\lambda = \tilde{A}_i^{-1/2} \tilde{\Sigma}_i$, and \tilde{A}_i is the diagonal elements of $\tilde{\Sigma}_i$. The sandwich working covariance structure $\tilde{\Sigma}_i = B_i^{1/2} R_i(\delta) B_i^{1/2}$ can be used to model the true $\tilde{\Sigma}_i = \text{Cov}(\varepsilon_i^2)$ with $B_i = 2\text{diag}\{\sigma_{i1}^4, \dots, \sigma_{in_i}^4\}$ and $R_i(\delta)$ mimics the correlation between ε_{ij}^2 and ε_{ik}^2 by introducing a new parameter δ . This idea was proposed by Ye and Pan (2006).

The parameter δ has little effect on the estimation in practice. Considering the AR(1) structure, we can estimate δ by the slope from the regression of $\log(\hat{\varepsilon}_{ij}^2, \hat{\varepsilon}_{ik}^2)$ on $\log(|t_j - t_k|)$. Details can be found in Example 4 in Liang and Zeger (1986). In our simulations and data analyses, the estimate of δ always falls in the interval $[0, 0.3]$. Moreover, results in Table 3 for simulation Study 2 also imply that we can ignore the difference between the independent and AR(1) structure for $R_i(\delta)$.

Penalized robust generalized estimating equations are distinguished from conventional generalized estimating equations in two aspects. First, the undesirable influence of outliers is controlled. In the core of the estimating equations, $\psi^\beta(\mu_i) = \psi(A_i^{-1/2}(y_i - \mu_i))$, $\psi^\gamma(\hat{r}_i) = \psi(D_i^{-1/2}(r_i - \hat{r}_i))$, and $\psi^\lambda(\sigma_i^2) = \psi(\tilde{A}_i^{-1/2}(\varepsilon_i^2 - \sigma_i^2))$. The function $\psi(\cdot)$ is chosen to limit the influence of outliers in the response variable, and a common choice is Huber's score function

$\psi_c(x) = \min\{c, \max\{-c, x\}\}$ for some constant c , $c = 2$ in our implementation. To ensure Fisher consistency, we use $C_i^\beta(\mu_i) = E[\psi(A_i^{-1/2}(y_i - \mu_i))]$, $C_i^\gamma(\hat{r}_i) = E[\psi(D_i^{-1/2}(r_i - \hat{r}_i))]$ and $C_i^\lambda(\sigma_i^2) = E[\psi(\tilde{A}_i^{-1/2}(\varepsilon_i^2 - \sigma_i^2))]$. When y_i 's are normal, the three expectations depend only on the choice of constant c in Huber's score function. We can also assign weights to each subject by diagonal weighting matrices $W_i^\beta = \text{diag}(w_{i1}^\beta, \dots, w_{in_i}^\beta)$, $W_i^\gamma = \text{diag}(w_{i1}^\gamma, \dots, w_{in_i}^\gamma)$ and $W_i^\lambda = \text{diag}(w_{i1}^\lambda, \dots, w_{in_i}^\lambda)$. Similar to Qin, Zhu, and Fung (2009), the weight function w_{ij} is chosen to be the Mahalanobis distance

$$w_{ij} = w(p_{ij}) = \min \left\{ 1, \left[\frac{b_0}{(p_{ij} - m_p)^T S_p^{-1} (p_{ij} - m_p)} \right]^{\rho/2} \right\},$$

with $\rho \geq 1$, where m_p and S_p are some robust estimates of the location and scale of p_{ij} such as the minimum covariance determinant estimators. We introduce the weight function to bound the influence of leverage points, covariate space only. As indicated in He, Fung, and Zhu (2005), we can include certain covariates that are likely to contribute to the leverage. In our simulation study, b_0 is chosen as the 95th percentile of the chi-squared distribution with degrees of freedom equal to the dimension of p_{ij} and ρ is fixed as 1. For simplicity and consistency, we choose $p_{ij} = x_{ij}$ for all three weighting matrices and denote them as $W_i = \text{diag}(w_{i1}, \dots, w_{in_i})$.

Selecting variables is achieved by adding a penalty term to each estimating equation. Usually, $q_{\tau(l)}(\cdot)$ is the first derivative for some penalty $p_{\tau(l)}(\cdot)$, where $l = 1, 2, 3$. For brevity, we replace $p_{\tau(1)}$, $p_{\tau(2)}$ and $p_{\tau(3)}$ by p_τ and $q_{\tau(1)}$, $q_{\tau(2)}$ and $q_{\tau(3)}$ by q_τ when no misunderstanding arises. In the simulation and data analyses, we only consider SCAD and ALASSO penalties to show the asymptotic properties. Fan and Li (2001) defined the smoothly clipped absolute deviation (SCAD) penalty function:

$$\begin{aligned} p_\tau(|\theta|) &= \tau(|\theta|)\{I((|\theta|) < \tau)\} + \frac{(a - |\theta|/2\tau)}{a - 1}I(\tau < (|\theta|) < a) \\ &\quad + \frac{a^2\tau}{2(a - 1)(|\theta|)}I\{(|\theta|) \geq a\tau\}, \end{aligned}$$

in which $a = 3.7$ was suggested by the authors. As a compromise between LASSO and Hard penalties, SCAD enjoys unbiasedness, sparsity and continuity properties simultaneously.

As a consistent version of the L_1 penalty, ALASSO penalty is $p_\tau = \tau|\theta|w$, for a known data-driven weight w . In this paper, we employ the weight $w = 1/|\hat{\theta}|$, where $\hat{\theta}$ is the regression coefficient estimate obtained from solving (2.2)–(2.4) without penalty.

2.3. Asymptotic properties

We write the m subjects-based penalized estimator $\hat{\theta}_m = ((\hat{\theta}_m^{s_1})^T, (\hat{\theta}_m^{s_2})^T)^T$ for the true value $\theta_0 = ((\theta_0^{s_1})^T, (\theta_0^{s_2})^T)^T$, where $\theta_0^{s_1}$ and $\theta_0^{s_2}$ are the nonzero and zero components of θ_0 , respectively. Denote the dimension of $\theta_0^{s_1}$ by s_1 and $s = s_1 + s_2$. The parameter space is assumed compact and the true value θ_0 is in the interior of the parameter space Θ .

In what follows we show that the penalized estimator $\hat{\theta}_m$ exists and converges to θ_0 at the rate $O_p(m^{-1/2})$, implying that it has the same consistency rate as the ordinary estimator. We prove that the \sqrt{m} -consistent estimator $\hat{\theta}_m^{s_1}$ is asymptotically normally distributed and possesses the oracle property under certain regularity conditions.

Theorem 1. *Under the assumptions in S1 of the supplementary material,*

- (a) *there exists an approximate zero-crossing solution $\hat{\theta}$ of $U(\theta) = 0$, such that $\hat{\theta} = \theta_0 + O_p(m^{-1/2})$;*
- (b) *for any \sqrt{m} consistent approximate zero-crossing solution of $U(\theta) = 0$, we have*

$$\lim_{m \rightarrow \infty} P\{\hat{\theta}_j = 0, j > s_1\} = 1.$$

The definition of the zero-crossing estimator $\hat{\theta}$, which is introduced in Johnson, Lin, and Zeng (2008), is given in S1 of the supplementary material. Theorem 1 implies that when we choose a proper τ_m , our robust penalized GEE approach can simultaneously achieve \sqrt{m} consistency of the regularized regression coefficient estimation and consistency of variable selection.

To obtain the asymptotic distribution of $\hat{\theta}$, we take $B = \lim_{m \rightarrow \infty} (1/m) \text{Cov}[U^R(\theta_0)]$ and assume it to be positive definite. The definition of $U^R(\theta_0)$ is given in supplement S1. We assume $\kappa_m(\theta) = E[\frac{1}{m} U^R(\theta)]$, $\kappa_m(\theta_0) = 0$, $\kappa_m(\theta)$ is continuous on Θ , and $\kappa_m(\theta)$ is differentiable at θ_0 with nonsingular derivative matrix G . Define $c_m = (q_{\tau_m}(|\theta_{01}^{s_1}|) \text{sgn}(\theta_{01}^{s_1}), \dots, q_{\tau_m}(|\theta_{0s_1}^{s_1}|) \text{sgn}(\theta_{0s_1}^{s_1}))^T$ and $\Omega = \text{diag}\{-q'_{\tau_m}(|\theta_0|) \text{sgn}(\theta_0)\}$, where τ_m is equal to either $\tau_m^{(1)}$, $\tau_m^{(2)}$, or $\tau_m^{(3)}$, depending on whether θ_{0j} is a component of β_0 , γ_0 , or λ_0 ($1 \leq j \leq s$). θ_{0j} is the j th component of θ_0 , and $\theta_{0j}^{s_1}$ is the j th component of $\theta_0^{s_1}$ ($1 \leq j \leq s_1$).

Theorem 2. *Under the conditions given in S1 of the supplementary material, for the SCAD penalty we have*

$$\sqrt{m}(G_m^{s_1} + \Omega_m^{s_1})\{\hat{\theta}_m^{s_1} - \theta_0^{s_1} + (G_m^{s_1} + \Omega_m^{s_1})^{-1}c_m\} \rightarrow N_{s_1}(0, B^{s_1})$$

in distribution, where B^{s_1} , $G_m^{s_1}$, and $\Omega_m^{s_1}$ are the $s_1 \times s_1$ submatrix of B , G , and Ω corresponding to the nonzero components $\theta_0^{s_1}$.

As a result, the asymptotic covariance matrix $\text{Cov}(\hat{\theta}_m^{s_1})$ of $\hat{\theta}_m^{s_1}$ is

$$\frac{1}{m}(G_m^{s_1} + \Omega_m^{s_1})^{-1} \hat{B}_m^{s_1} (G_m^{s_1} + \Omega_m^{s_1})^{-1}.$$

Thus, our proposed robust penalized joint model based on the SCAD penalty possesses the oracle property. Proofs of Theorems are sketched in S1 of the supplementary material.

2.4. Implementation

An iterative MM algorithm for estimating β , γ and λ is described in detail in S2 of the supplementary material. Meanwhile, the choice of τ is critical. In practice, we select $\tau^{(1)}$ by minimizing the robustified generalized cross-validation (GCV) criterion

$$\text{GCV}_\beta(\tau) = \frac{RSS_\beta(\tau)/m}{\{1 - d(\tau)/m\}^2},$$

where $RSS_\beta(\tau)$ is the robustified residual sum of squares

$$\sum_{i=1}^m [W_i \psi(A_i^{-1/2}(y_i - x_i \hat{\beta}_\tau))]^2,$$

and $d(\tau)$ is the effective number of parameters, $d_\beta(\tau) = \text{tr}([\hat{G}(\beta) + \Delta_\tau(\hat{\beta}_\tau)]^{-1} [\hat{G}(\beta)]^T)$; here $\hat{\beta}_\tau$ is the solution of the penalized robust GEE when τ is fixed. We select $\hat{\tau} = \text{argmin}_\tau \text{GCV}_\beta$.

Similar to the choice of $\tau^{(1)}$ we select tuning parameters $\tau^{(2)}$ and $\tau^{(3)}$ by minimizing the robustified GCV statistics

$$\text{GCV}_\gamma(\tau) = \frac{RSS_\gamma(\tau)/m}{\{1 - d_\gamma(\tau)/m\}^2}, \quad \text{GCV}_\lambda(\tau) = \frac{RSS_\lambda(\tau)/m}{\{1 - d_\lambda(\tau)/m\}^2},$$

where $RSS_\gamma(\tau) = \sum_{i=1}^m [W_i \psi(D_i^{-1/2}(r_i - \hat{r}_i))]^2$ and $RSS_\lambda(\tau) = \sum_{i=1}^m [W_i \psi(\tilde{A}_i^{-1/2}(\hat{\varepsilon}_i^2 - \hat{\sigma}_i^2))]^2$. $d_\gamma(\tau)$ and $d_\lambda(\tau)$ are the effective numbers of covariance parameters.

To avoid computational burden, we recommend selecting parameters sequentially, as follows

- (1) Fix $\tau^{(2)} = \tau^{(3)} = 0$, choose $\hat{\tau}^{(1)} = \text{argmin}_{\tau^{(1)}} \text{GCV}_\beta(\tau^{(1)})$;
- (2) Fix $\tau^{(1)} = \hat{\tau}^{(1)}$ and $\tau^{(3)} = 0$, choose $\hat{\tau}^{(2)} = \text{argmin}_{\tau^{(2)}} \text{GCV}_\gamma(\tau^{(2)})$;
- (3) Fix $\tau^{(1)} = \hat{\tau}^{(1)}$ and $\tau^{(2)} = \hat{\tau}^{(2)}$, choose $\hat{\tau}^{(3)} = \text{argmin}_{\tau^{(3)}} \text{GCV}_\lambda(\tau^{(3)})$;
- (4) The final choice is $(\hat{\tau}^{(1)}, \hat{\tau}^{(2)}, \hat{\tau}^{(3)})$.

The process of tuning parameter selection has merits aside from reducing the computational burden. From a numerical point of view, minimizing GCV_γ and

GCV_λ does not always benefit mean estimation. Sequentially selecting tuning parameters lets us adjust and stabilize selection while balancing the importance of estimation for the mean and the covariance structure.

3. Simulation

We conducted a simulation study to assess the performance of the proposed estimators on three aspects: The efficiency of our robust model compared with the corresponding non-robust version; the necessity of the robust method for the presence of outliers; comparison with the classical GEE method under covariance matrix misspecification.

We compared model errors of different variable selection procedures using median of model error (MME), where model errors are evaluated, following Fan and Li (2001) as

$$ME_\beta = (\hat{\beta} - \beta_0)^T X X^T (\hat{\beta} - \beta_0),$$

$$ME_\gamma = (\hat{\gamma} - \gamma_0)^T T T^T (\hat{\gamma} - \gamma_0), \quad ME_\lambda = (\hat{\lambda} - \lambda_0)^T Z Z^T (\hat{\lambda} - \lambda_0),$$

where $X = (X_1^T, \dots, X_m^T)^T$, $T = (T_1^T, \dots, T_m^T)^T$, and $Z = (Z_1^T, \dots, Z_m^T)^T$. We employ average correctly fit percentage (CF%) to measure the accuracy of the model selection procedure, where correctly fit means that the procedure selects the exact subset model. Moreover, we compare the average numbers of regression coefficients that are correctly shrunk (CS) to zeros. Replications of each scenario were at 200.

We generated balanced data sets, $n_i = n$, for convenience. In practice, our method also works well for the nearly balanced data set (Zhou and Qu (2012)). In the data example, we can also handle unbalanced data when the observation time information is available; this is a reasonable assumption for longitudinal data, as the subjects' measurements are recorded along with the observation time.

Study 1. We simulated 100 (or 200) subjects, each of which had 5 observations that were multivariate normal $N_5(\mu_i, \Sigma_i)$. The true values of the mean parameter and log-innovation variances were $\beta = (3, 0, 0, -2, 1, 0, 0, 0, 0, -4)^T$ and $\lambda = (0, 1, 0, 0, 0, -2, 0)^T$, respectively. Two specifications are designed for generalized autoregressive parameters; $\gamma = (0, 0, 0, 0, 0, 0, 0)^T$ and $\gamma = (0.2, 0, 0, 0, 0, 0, 0)^T$. The mean covariates $x_{ij} = (x_{ijt})_{t=1}^{10}$ were drawn from the multivariate normal with mean 0 and covariance matrix of AR(1) structure with variance $\sigma^2 = 1$ and correlation parameter $\rho = 0.5$ ($i = 1, \dots, 100$; $j = 1, \dots, 5$). Then $g_{ijk} = (x_{ijt} - x_{ikt})_{t=1}^7$ and $z_{ij} = (x_{ijt})_{t=1}^7$ were covariates for the generalized autoregressive parameters and the log-innovation variances. Using these values, the mean μ_i and covariance matrix Σ_i were constructed through the modified Cholesky decomposition from (2.2)–(2.4). The responses y_i 's were then drawn from $N(\mu_i, \Sigma_i)$ ($i = 1, \dots, 100$).

We consider two contaminations:

C0: randomly choose 1% of x_{ij1} to be $x_{ij1} - 1$;

C1: randomly choose 2% of x_{ij1} to be $x_{ij1} - 3$ and 2% of y_{ij} to be $y_{ij} + 10$.

NC represents no contamination situation hereafter. C1 is a commonly used contamination setting in previous research on robust methods, and we consider a tiny contamination in C0. The initial value of the mean estimation was obtained from the robust GEE estimation in He, Fung, and Zhu (2005) with independent working correlation that guarantees consistency of the autoregressive parameters and innovative parameters after the first iteration. If the covariance matrix falls into the spanned space of the covariates, the proposed method converges quickly under no contamination, usually in a few steps. However, in non-robust modeling, the traditional MM algorithm has a large probability of non-convergence under contaminations. Typically, in C1, 180 of 200 replications cannot obtain a final estimation due to divergence of iteration in covariance estimation, even when the initial value of estimation for the mean parameter β is close to the true value. Moreover, in 10%-20% of replications the non-robust algorithm fails to converge in C0, where the perturbation is almost negligible.

To check asymptotic properties, we simulated with sample size of 100 and 200 subjects. In Table 1, we list the median of model error (MME), average correctly fit percentage (CF%), and average numbers of regression coefficients that are correctly shrunk (CS) for both robust and non-robust methods, under NC (no contamination) and C0 (tiny contamination). Notice that the results for C0 were obtained based on convergence cases only. R_{scad} and R_{alasso} represent the proposed robust method employing SCAD and ALASSO, respectively. NR is the non-robust method.

As the number of subjects m increases, the MME of both robust and non-robust methods decreases, while CF% and CS approach 1 and the true number of zero parameters, respectively. These are consistent with the oracle property. The R_{scad} and NR methods perform equally well in variable selection under NC, although an acceptable loss of efficiency for the robust method can be detected from the slightly larger MME (in NC R_{scad} compared to NC NR). However, under C0, the robust method apparently outperforms the non-robust method in both estimation efficiency and variable selection. Especially in covariance model identification, the non-robust method fails to correctly identify the innovation variance model under small contamination in most replications.

We compare the performance of R_{scad} and R_{alasso} to find that R_{scad} outperforms R_{alasso} in β and λ estimation and R_{alasso} performs better in γ estimation. In fact, SCAD allows almost no penalty if the true parameter is far from 0 while ALASSO penalizes all parameters, which increases the MME, especially in λ . In

Table 1. Parameter estimation and selection in Study 1.

	$\gamma = 0$						$\gamma \neq 0$					
	$n = 100$			$n = 200$			$n = 100$			$n = 200$		
	MME	CF%	CS	MME	CF%	CS	MME	CF%	CS	MME	CF%	CS
NC NR _{scad}												
β	0.040	66.0	5.61	0.030	80.5	5.79	0.032	56.5	5.47	0.021	73.5	5.70
γ	0.016	36.5	5.86	0.000	76.0	6.62	0.021	45.0	5.09	0.004	86.0	5.82
λ	0.057	73.5	4.71	0.032	93.5	4.93	0.043	77.5	4.71	0.026	96.0	4.96
NC R _{scad}												
β	0.050	72.0	5.68	0.040	84.0	5.82	0.039	53.0	5.38	0.026	73.0	5.70
γ	0.018	35.5	5.77	0.000	69.5	6.49	0.025	37.5	5.00	0.006	76.0	5.67
λ	0.078	92.5	4.92	0.048	100	5.00	0.060	93.0	4.92	0.040	99.5	5.00
NC R _{alasso}												
β	0.047	54.5	5.41	0.033	75.5	5.73	0.049	52.0	5.38	0.030	69.0	5.65
γ	0.012	44.5	6.01	0.000	72.0	6.50	0.021	47.0	5.00	0.003	85.5	5.85
λ	0.242	51.0	4.28	0.193	73.5	4.71	0.318	67.0	4.57	0.265	86.0	4.84
C0 NR _{scad}												
β	0.068	65.0	5.57	0.067	81.5	5.79	0.068	57.5	5.42	0.066	71.5	5.63
γ	0.021	33.0	5.66	0.000	59.5	6.37	0.061	28.0	4.45	0.039	52.5	5.14
λ	0.331	17.5	3.12	0.520	15.0	3.02	0.408	19.0	3.06	0.575	12.0	2.83
C0 R _{scad}												
β	0.053	72.0	5.67	0.043	85.0	5.84	0.052	52.0	5.34	0.065	87.5	5.87
γ	0.018	35.0	5.82	0.000	72.0	6.55	0.040	31.5	4.77	0.000	71.0	6.53
λ	0.068	91.0	4.91	0.052	99.5	5.00	0.088	85.0	4.83	0.069	99.0	4.99
C0 R _{alasso}												
β	0.055	52.5	5.41	0.038	74.0	5.70	0.091	50.5	5.29	0.061	64.5	5.54
γ	0.013	41.5	5.93	0.000	74.5	6.55	0.049	28.5	4.54	0.017	71.5	5.61
λ	0.280	43.5	4.22	0.224	67.5	4.63	0.550	60.0	4.49	0.420	84.0	4.82

Simulation results of median of model error (MME), average correctly fit percentage (CF%), and average numbers of regression coefficients that are correctly shrunk (CS) for both robust (R_{scad}, R_{alasso}) and non-robust (NR) methods, under NC (no contamination) and C0 (tiny contamination), with 200 replications.

sum, although both robust methods can resist the contamination according to our simulations, R_{scad} is preferred.

Standard errors for SCAD estimators in Study 1 for non-zero parameters (β_1 , β_4 , β_5 , β_{10} , γ_1 , λ_2 , and λ_6) are attached in Table S3.1 of the supplementary material. In fact, the standard errors of the robust method are close to those of the non-robust method under no contamination and are much smaller under contamination C0. To investigate the influence of outliers in covariance estimation, we list entropy losses and quadratic losses (defined in S3 of the supplementary material) in Table 2. Again, we find poor performance of the non-robust method in C0 compared to the robust approach.

Table 2. Entropy loss (L_1) and quadratic loss (L_2) in estimating Σ in Study 1.

		$\gamma = 0$				$\gamma \neq 0$			
		$n = 100$		$n = 200$		$n = 100$		$n = 200$	
		L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2
NC	NR	0.097	0.253	0.024	0.091	0.104	0.249	0.038	0.097
	R_{scad}	0.117	0.318	0.036	0.129	0.135	0.361	0.054	0.144
	R_{lasso}	0.189	0.756	0.122	0.528	0.251	1.112	0.151	0.698
C0	NR	0.255	1.070	0.289	1.454	0.463	2.318	0.479	2.911
	R_{scad}	0.107	0.272	0.038	0.139	0.190	0.601	0.049	0.187
	R_{lasso}	0.215	0.900	0.134	0.610	0.436	2.289	0.286	1.473

Study 2. In this study, we looked into the effect of covariance misspecification on mean estimation. We compared the performance of the robust penalized joint model, denoted rpj in Table 3, method with that of robust and non-robust penalized generalized estimating equations, denoted $rpgee$ and $pgee$, methods that assume a fixed working correlation matrix and solve (2.2) for the mean. The most salient difference of the three methods is that our joint model builds regression models after decomposing the covariance matrix, while $RPGEE$ and $PGEE$ treat the covariance matrix as a nuisance and the marginal variance of y_i is estimated from the sample. In $PGEE$, we set the tuning parameter of the Huber function c as 1,000 and the weight $W = I$ when solving (2.2) for the mean.

Under the same mean set-up as in Study 1, we considered the covariance structures working independence (IN), auto-regressive (AR), and exchangeable (EX) with correlation parameter 0.5. We compared the performance of eight estimators: rpj_{ar} and rpj_{in} are robust joint models with independent and AR(1) correlation structure assumptions for $Cov(\varepsilon_i^2)$ in (2.4), respectively; $rpgee_{in}$, $rpgee_{ar}$, and $rpgee_{ex}$ (or $pgee_{in}$, $pgee_{ar}$, and $pgee_{ex}$) are robust (or non-robust) penalized GEE estimations with IN, AR and EX as working correlation matrices. We adopted SCAD penalty in the study.

Table 3 lists the results under NC and C1. In the table, we employed MRME (the median of relative model error) to compare the performance of the eight estimators

$$\frac{\text{ME (model error) of the estimator}}{\text{ME of PGEE estimator with true covariance matrix}}.$$

In the absence of outliers (NC), in general, $pgee$ estimators performed slightly better than $rpgee$ estimators which in turn were better than the rpj estimators. Besides, rpj_{in} and rpj_{ar} had similar performances, which suggests that the choice of the δ in (2.4) has little effect on the mean and covariance estimation. As a result, we fixed $\delta = 0$ in a later application.

Table 3. Simulation results for Study 2.

		IN			AR			EX		
		MRME	CF%	CS	MRME	CF%	CS	MRME	CF%	CS
NC	rpj _{ar}	1.09	92.0	5.92	1.33	76.0	5.74	1.87	67.5	5.63
	rpj _{in}	1.09	92.0	5.92	1.34	75.5	5.73	1.87	67.5	5.63
	rpgee _{ar}	1.04	94.5	5.94	1.12	96.0	5.96	1.56	92.0	5.92
	rpgee _{ex}	1.04	93.5	5.93	1.20	92.0	5.92	1.15	91.0	5.91
	rpgee _{in}	1.07	93.0	5.93	1.33	80.0	5.79	1.92	69.0	5.65
	pgee _{ar}	0.99	91.5	5.91	1.01	94.5	5.95	1.41	89.5	5.90
	pgee _{ex}	0.99	90.5	5.90	1.11	86.5	5.87	1.01	83.5	5.83
	pgee _{in}	1.00	91.5	5.92	1.26	72.0	5.70	1.83	63.0	5.57
C1	rpj _{ar}	0.08	92.0	5.92	0.08	95.5	5.96	0.10	89.0	5.89
	rpj _{in}	0.08	92.0	5.92	0.08	95.5	5.96	0.10	89.0	5.89
	rpgee _{ar}	0.16	90.0	5.90	0.11	88.5	5.88	0.13	83.0	5.83
	rpgee _{ex}	0.16	89.5	5.90	0.10	88.5	5.88	0.13	88.0	5.88
	rpgee _{in}	0.15	89.0	5.89	0.09	76.5	5.75	0.11	71.0	5.68
	pgee _{ar}	0.95	30.0	4.94	0.62	67.0	5.63	0.76	62.5	5.59
	pgee _{ex}	0.96	30.5	4.93	0.58	65.0	5.58	0.75	67.5	5.62
	pgee _{in}	0.92	28.0	4.88	0.52	54.0	5.43	0.63	48.0	5.34

Simulation results of median of relative model error (MRME), average correctly fit percentage (CF%), and average numbers of regression coefficients that are correctly shrunk (CS) under NC (no contamination) and C1 (contamination), with 200 replications. rpj_{ar} and rpj_{in} are robust joint models with independent and AR(1) correlation structure in (2.4), respectively; rpgee_{in}, rpgee_{ar}, and rpgee_{ex} (or pgee_{in}, pgee_{ar}, and pgee_{ex}) are robust (or non-robust) penalized GEE estimations with IN, AR, and EX as working correlation matrices.

Under C1, rpj estimators improved substantially over the rpgee and pgee estimators. Without the bounded score on the mean estimator, pgee mean estimators collapsed in any of the IN, AR or EX covariance structure, as all MRME's for rpj and rpgee are less than 1 in C1. By adopting the robust estimator for the mean, rpgee estimators had reasonable performances on the mean estimation. However, rpj estimators further improved the performance of the mean estimator (variable selector). Standard errors for the mean estimators can be found in supplementary Table S3.2, which reveals the influence of outliers on the mean estimation.

Table S3.3 of the supplementary material lists entropy losses and quadratic losses on covariance matrix estimation. When there is no contamination, rpgee and pgee had comparable performances, and better than rpj. However, under C1, rpgee and pgee can be seriously affected by the contamination.

4. Data Analysis

In this section, we illustrate our method analyzing the hormone data that

has been analyzed by Fung et al. (2002), He, Zhu, and Fung (2002), Fan, Qin, and Zhu (2012), and Qin, Zhu, and Fung (2009). The data set contains 492 observations of progesterone level within a menstrual cycle, collected from 34 women clinical participants. In our model, the response variable y_{ij} is the log-transform of progesterone level and the observation time is t_{ij} , apart from which patient's age and body mass index (BMI) are recorded. Two objectives were considered when we implemented the robust variable selection method in the joint mean and covariance model to this hormone data set: to accommodate the influence of outliers and leverage points, and to detect outliers in the data set; to identify statistically significant covariates in linear models of both mean and covariance matrix.

The starting mean model we proposed was

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 Age_i + \beta_2 BMI_i + \beta_3 t_{ij} + \beta_4 t_{ij}^2 \\ &\quad + \beta_5 Age_i \times BMI_i + \beta_6 Age_i \times t_{ij} + \beta_7 BMI_i \times t_{ij} + e_{ij} \\ &= x_{ij}^T \beta + e_{ij}. \end{aligned}$$

For the covariance model, we followed the model in (1) and chose the corresponding covariates as

$$g_{ijk} = (1, (t_{ij} - t_{ik}), (t_{ij} - t_{ik})^2, (t_{ij} - t_{ik})^3)^T, \quad z_{ij} = x_{ij}.$$

Three estimators were under consideration: rpj is the robust penalized joint model; pj is the penalized joint model proposed by Kou and Pan (2011); gee is the widely-used GEE estimator. Table 4 summarizes estimators of the mean parameters with standard errors. We noticed that the joint models (rpj and pj) are more parsimonious than gee as they choose time as the only significant variable. This is consistent with previous research that found that both Age, BMI, their interaction and interactions with time are not statistically significant in the model. We found that the regression coefficients for Time obtained by rpj and pj rather different, the non-robust pj estimator is affected by outliers.

Estimates with standard errors for the generalized autoregressive coefficients and innovation variances are summarized in Table 5. We found that the cubic polynomial of time is statistically significant for autoregressive coefficients γ in all fits. Standard errors in non-robust method are larger than those in the robust method. Unlike estimators of generalized autoregressive coefficients, significant covariates for innovation variances were not found in our analysis. Due to the existence of outliers, the non-robust method failed to select significant covariates for innovation variances.

Outlier detection was practiced with our procedure. By investigating the standardized residuals s_{ij} and the weight function w_{ij} , we found one observation-level outlier (observation 10) and one subject-level leverage point (subject 18).

Table 4. Estimators of the mean parameters β and standard errors (inside brackets) for progesterone data.

	Intercept	Age	BMI	Time	Time ²	Age \times BMI	Age \times Time	BMI \times Time
rpj	0.837 (0.074)	0 (-)	0 (-)	0.691 (0.053)	0 (-)	0 (-)	0 (-)	0 (-)
pj	0.892 (0.078)	0 (-)	0 (-)	0.562 (0.056)	0 (-)	0 (-)	0 (-)	0 (-)
gee	0.870 (0.126)	1.684 (2.180)	-2.671 (2.928)	0.709 (0.049)	0.186 (0.085)	-4.829 (50.09)	1.493 (0.827)	0.701 (0.857)

In the table, rpj represents the robust penalized joint model; pj represents penalized joint model proposed by Kou and Pan (2011); gee is the traditional GEE estimator. Working independence is considered.

Table 5. Estimates of the generalized autoregressive parameters γ and innovation variance parameters λ for progesterone data.

	γ_1	γ_2	γ_3	γ_4				
rpj	0.902 (0.056)	-2.726 (0.203)	2.339 (0.190)	-0.579 (0.050)				
pj	0.882 (0.063)	-2.623 (0.234)	2.185 (0.223)	-0.523 (0.060)				
	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
rpj	-1.111 (0.103)	0 (-)	0 (-)	0 (-)	0.180 (0.120)	45.30 (34.50)	0 (-)	-1.317 (2.435)
pj	-0.882 (0.099)	-1.527 (1.398)	-0.283 (1.761)	0.186 (0.086)	0.029 (0.115)	56.98 (29.03)	0.175 (1.699)	1.313 (2.044)

$p_{ij} = (AGE_i, BMI_i)$ contributed to the weight functions w_{ij} in our robust method. Subject 18 is a leverage point which had not been identified before; it has an extremely high BMI of 38 that heavily downweights the cluster of observations from the patient. A careful inspection of the standardized residual s_{ij} tells us that case 10 is the most extreme point with $s_{ij} = -6.09$. The progesterone level of the 10th observation for subject 1 (case 10) is 2.46, which is very different from its neighborhood observations 9 and 11 measured one day before and one day after. This inconsistency has not been noticed in the literature. All other thirteen observations of subject 1 range from 8.5 to 13.4 and in particular, this observation was the lowest progesterone level in the data set. We concluded that case 10 of subject 1 is a clear outlier. The influence of this outlier on the parameter estimates is detailed in Table S4.1 of the supplementary material. We also noticed that some observations have large standardized residuals, such as cases 117, 334 and 372, due to the fact that they are extreme values of the progesterone level within a subject. The effects on these potential outliers are downweighted by our robust method in the estimation of mean and covariance

parameters.

Acknowledgement

The authors are grateful to two reviewers, an associate editor, and the Co-Editor for their insightful comments and suggestions which have improved the manuscript significantly. This work was supported by the National Natural Science Foundation of China (10931002, 11271180).

References

- Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.
- Cantoni, E., Flemming, J. M. and Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics* **61**, 507-514.
- Cook, R. D and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1-10.
- El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* **36**, 2717-2756.
- Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **35**, 632-641.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.
- Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *J. Amer. Statist. Assoc.* **103**, 1520-1533.
- Fan, Y. L., Qin, G. Y. and Zhu, Z. Y. (2012). Variable selection in robust regression models for longitudinal data. *J. Multivariate Anal.* **109**, 156-167.
- Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59**, 126-132.
- Fung, W. K., Zhu, Z. Y., Wei, B., and He, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *J. Roy. Statist. Soc. Ser. B* **64**, 565-579.
- He, X., Fung, W. K., and Zhu, Z. Y. (2005). Robust estimation in generalized partial linear models for clustered data. *J. Amer. Statist. Assoc.* **472**, 1176-1184.
- He, X., Zhu, Z. Y., and Fung, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579-590.
- Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85-98.
- Jeng, X. J. and Daye, Z. J. (2011). Sparse covariance thresholding for high-dimensional variable selection. *Statist. Sinica* **21**, 625-657.
- Johnson, B., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *J. Amer. Statist. Assoc.* **103**, 672-680.
- Kou, C. and Pan, J. (2011). Variable selection for joint mean and covariance models via penalised Likelihood, Technical report, Probability and Statistics Group School of Mathematics, The University of Manchester.
- Leng, C., Zhang, W., and Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *J. Amer. Statist. Assoc.* **105**, 181-193.

- Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- Mao, J., Fung, W. K. and Zhu, Z. Y. (2011). Joint estimation of mean-covariance model for longitudinal data with basis function approximations. *Comput. Statist. Data Anal.* **55**, 983-992.
- Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **73**, 13-22.
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* **86**, 677-690.
- Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425-435.
- Qin, G. Y., Zhu, Z. Y., and Fung, W. K. (2009). Robust estimation of covariance parameters in partial linear model for longitudinal data. *J. Statist. Plann. Inference* **139**, 558-570.
- Qu, A., Lindsay, B., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823-836.
- Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104**, 177-186.
- Xu, J. and Mackenzie, G. (2012). Modelling covariance structure in bivariate marginal models for longitudinal data. *Biometrika* **99**, 649-662.
- Ye, H. and Pan, J. (2006). Modelling covariance structures in generalized estimating equations for longitudinal data. *Biometrika* **93**, 927-941.
- Zheng, X. Y., Fung, W. K., and Zhu, Z. Y. (2013). Robust estimation in joint mean-covariance regression model for longitudinal data. *Ann. Inst. Statist. Math.* **65**, 617-638.
- Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *J. Amer. Statist. Assoc.* **107**, 701-710.
- Zou, H. (2006). The Adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Department of Biostatistics, Fudan University, Shanghai, China.

E-mail: xyzheng@fudan.edu.cn

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong.

E-mail: wingfung@hku.hk

Department of Statistics, Fudan University, Shanghai, China.

E-mail: zhuzy@fudan.edu.cn

(Received October 2011; accepted February 2013)